ORIGINAL ARTICLE: Clinical Endoscopy

Development of a deep learning model for the histologic diagnosis of dysplasia in Barrett's esophagus

Shahriar Faghani, MD,[1,2,*] D. Chamil Codipilly, MD,[3,*] David Vogelsang, MD,[1,2] Mana Moassefi, MD,[1,2] Pouria Rouzrokh, MD, MPH,[1,2] Bardia Khosravi, MD, MPH,[1,2] Siddharth Agarwal, MBBS,[3] Lovekirat Dhaliwal, MBBS,[3] David A. Katzka, MD,[3] Catherine Hagen, MD,[4] Jason Lewis, MD,[5] Cadman L. Leggett, MD,[3] Bradley J. Erickson, MD, PhD,[1,2,*] Prasad G. Iyer, MD[3,*]

Rochester, Minnesota; Jacksonville, Florida, USA

Background and Aims: The risk of progression in Barrett's esophagus (BE) increases with development of dysplasia. There is a critical need to improve the diagnosis of BE dysplasia, given substantial interobserver disagreement among expert pathologists and overdiagnosis of dysplasia by community pathologists. We developed a deep learning model to predict dysplasia grade on whole-slide imaging.

Methods: We digitized nondysplastic BE (NDBE), low-grade dysplasia (LGD), and high-grade dysplasia (HGD) histology slides. Two expert pathologists confirmed all histology and digitally annotated areas of dysplasia. Training, validation, and test sets were created (by a random 70/20/10 split). We used an ensemble approach combining a "you only look once" model to identify regions of interest and histology class (NDBE, LGD, or HGD) followed by a ResNet101 model pretrained on ImageNet applied to the regions of interest. Diagnostic performance was determined for the whole slide.

Results: We included slides from 542 patients (164 NDBE, 226 LGD, and 152 HGD) yielding 8596 bounding boxes in the training set, 1946 bounding boxes in the validation set, and 840 boxes in the test set. When the ensemble model was used, sensitivity and specificity for LGD was 81.3% and 100%, respectively, and >90% for NDBE and HGD. The overall positive predictive value and sensitivity metric (calculated as F1 score) was .91 for NDBE, .90 for LGD, and 1.0 for HGD.

Conclusions: We successfully trained and validated a deep learning model to accurately identify dysplasia on whole-slide images. This model can potentially help improve the histologic diagnosis of BE dysplasia and the appropriate application of endoscopic therapy. (Gastrointest Endosc 2022;96:918-25.)

(footnotes appear on last page of article)

Barrett's esophagus (BE) is the only known precursor lesion for esophageal adenocarcinoma (EAC), a malignancy with a 5-year survival of <20%.[1] The risk of malignant progression in BE increases with the development of low-grade dysplasia (LGD) and high-grade dysplasia (HGD).[2] Gastroenterology societies recommend surveillance to detect dysplasia and initiate endoscopic eradication therapy (EET) for those with LGD and HGD, because this has been proven to decrease progression to EAC.[3,4]

Unfortunately, the histologic diagnosis of dysplasia, particularly LGD, carries significant challenges. For LGD, the subtlety of pathologic changes and the visual similarities of postinflammatory regenerative changes likely contribute to this diagnostic challenge.[5] This leads to substantial interobserver variability among pathologists when interpreting LGD. Furthermore, LGD is overdiagnosed in the community and is frequently downstaged to NDBE by experienced GI pathologists.[6] The distinction is clinically important given the recommendation for considering EET in LGD and substantially lower progression rates of NDBE compared with LGD. As such, there is a critical need to improve diagnostic capabilities for BE dysplasia. Such a tool could avoid the need for slide review by multiple pathologists, assuage unnecessary patient concern, and more accurately help assign endoscopic surveillance and ablative strategies.

Deep learning algorithms can be used for visual object recognition.[7,8] The use of convolutional neural networks has been reported in the real-time endoscopic diagnosis of BE dysplasia and other GI pathologies.[9,10] A study analyzing 180 BE whole-slide images (WSIs) demonstrated modest accuracy in diagnosing BE dysplasia with a deep learning attention–based neural network.[11] Another deep learning study that used 57 WSIs and mass spectrometry demonstrated good performance in distinguishing epithelial from stromal tissue but modest performance on tile-level diagnosis of BE.[12]

Considering the tiny fraction of a whole slide that may harbor dysplasia,[13] use of a convolutional neural network alone can lead to considerable inefficiency because of computational costs and time to analyze the millions of pixels that make up a slide. However, regions of interest identified with an object detection model, which performs efficient first-pass prediction, can then be assessed with an independent classifier model, mitigating the need for excessive computational costs and improving performance overall. Further, WSI-level diagnosis based on a combination of models can be more accurate because of the ensemble theorem, which states that multiple independent classifiers offer better performance than a single classifier.[13]

We aimed to develop a fully automated, deep learning model to distinguish dysplasia grade (NDBE, LGD, and HGD) on WSIs from a large cohort of digitized BE histology slides. Given recommendations for EET for either LGD or HGD, our secondary aim was to assess model accuracy in distinguishing a composite finding of any degree of dysplasia (LGD or HGD) from NDBE.

METHODS

The Mayo Clinic institutional review board approved this study.

Slide selection and digitization

BE cases (NDBE, LGD, and HGD) from January 1992 to September 2020 were identified by performing an electronic search of the Mayo Clinic electronic health record using appropriate International Classification of Diseases, Ninth Edition and Tenth Edition codes for BE (530.85 and K22.7, respectively). We included records from 1992 onward because all endoscopy and pathology reports from that year are electronically accessible. A spreadsheet of potential patients within each class was obtained, and each patient was given a random number using the "Rand ()" function on Microsoft Excel (Microsoft Corp, Redmond, Wash, USA). These randomly generated numbers were then sorted from least to highest to randomize the list of patients within each pathology class. All BE patient histology slides were first reported by a GI pathologist with expertise in BE dysplasia diagnosis.

The BE diagnosis was confirmed by a review of the endoscopic and histologic reports. To be included in this study, patients had to provide authorization for inclusion into research studies, have endoscopic evidence of BE ($\geq$1 cm of columnar mucosa extending proximally from the gastroesophageal junction), and a confirmed histologic diagnosis of BE (intestinal type mucosa with goblet cells). One histology slide with the highest level of dysplasia per patient was chosen for this study. Patients with a prior history of ablative therapies were excluded from the present study, but patients with a history of (or concurrent) endoscopic resection were eligible for inclusion.

Hematoxylin and eosin slides from the randomized patient list were re-reviewed by study pathologists (C.H. and J.L.) with significant experience in diagnosing BE-related dysplasia to confirm the degree of dysplasia based on established criteria.[14,15] Briefly, slides were examined at scanning magnification, looking for areas of increased nuclear staining (hyperchromasia) or architectural complexity. If either were identified, then higher power evaluation was performed. Cases with intestinal metaplasia and no cytologic or architectural abnormalities or those in which only mild hyperchromasia was present at the basal, proliferative zones with maturation at the surface were classified as NDBE. LGD was diagnosed when there was an increase in nuclear size and hyperchromasia with nuclear crowding, typically involving the surface epithelium, but no significant loss of polarity with respect to the basement membrane, macronucleoli, or architectural abnormalities present. HGD was defined by glands with larger nuclei than those seen in LGD, along with significant variability in nuclear size and shape (pleomorphism), loss of polarity, greater mitotic activity, and architectural changes such as budding, branching, and cribriform formation.

Because the initial clinical read was made by an expert GI pathologist, if 1 study pathologist agreed with the review at the time of the present study, the slide was included. If there was disagreement, the slide was reviewed by the other study pathologist, and if both study pathologists agreed on the pathology, the slide was included. However, if there was disagreement between study pathologists, the slide was returned, and a new slide was called for assessment of inclusion into the study. Confirmed slides were then digitized using a high-resolution pathology slide scanner (Aperio AT2; Leica Biosystems Inc, Buffalo Grove, Ill, USA).

Once digitized, study pathologists (C.H. and J.L.) again reviewed slides and digitally annotated areas with the highest grade of dysplasia in each slide using Aperio ImageScope software (Leica Biosystems Inc) (Fig. 1A). Glands of interest were identified and annotated as precisely as possible, taking care to not include adjacent glands that were not representative of the pathology being evaluated in each case. We aimed to annotate at least 5 areas of the dysplasia grade of interest per slide. Annotations were saved as extensible markup language (.XML) files. The methodology for preparation of slides for analysis by our deep learning algorithm is presented in the Supplementary Methods (available online at www.giejournal.org).

Training, validation, and testing sets

Pathology annotations were randomly split by patient into training, validation, and test sets in a 70%/20%/10% split using the Scikit-learn library[16] before any training or hyperparameter tuning. The training set was used to train the model to recognize the specific characteristics that differentiate the 3 histology classes (NDBE, LGD, and HGD). The validation set allowed optimizing hyperparameters during training, but results were not directly used to update parameters. The test set was used for pure, unbiased assessment of slide results without feedback or alteration to model training or hyperparameter tuning. Pathologist annotations were removed from test set slides before WSI analysis. The ratio of NDBE/LGD/HGD in the slide dataset was 1:1.5:1, the training set was balanced by oversampling the HGD and NDBE in training data, and the same training set was used for both the object detection and classifier models to avoid data leakage.

Model development

We developed an ensemble approach using 2 distinct deep learning models to sequentially analyze WSIs.[17] We created an initial object detection model to identify pathologic regions of interest and perform first-pass dysplasia grade prediction on WSIs using a you only look once (YOLO) v5 model (Fig. 1B-E).[18,19] Subsequently, a second convolutional neural network–based classifier model was used to increase classification accuracy within the identified regions of interest (Fig. 1F-H).[20-22] The specific development of these models is further explained in the Supplementary Methods. We developed the models using Pytorch 1.7.1, Fastai 2.0.0 on Python 3.7 on a GPU cluster of 4 GPUs (NVIDIA A100, NVIDIA Corporation, Santa Clara, Calif, USA).

WSI-level prediction using an ensemble approach

We converted each WSI into nonoverlapping 1280 × 1280-pixel tiles. For the test set, annotations from pathology review were removed to allow unbiased assessment. These were fed into the object detection (YOLOv5) model (Fig. 1I and J), and detected regions were then resized into 224 × 224-pixel tiles for input into the classifier model (Fig. 1K). Because the YOLO model also produces a class label for the region, 2 predictions were made in each area of interest (Fig. 1K and L); the results for each WSI were organized in a 3 × 3 matrix (Fig. 2). The first column demonstrates the object detection (YOLO) predictions, and the first row displays the classifier model's predictions. In the matrix, the diagonal axis shows the agreement between the object detection and classifier models. We chose the highest grade of dysplasia (or presence or absence of dysplasia in the 2-class model) as agreed on by both models as the final WSI histologic prediction.

RESULTS

Patients included in the cohort

We initially identified a list of 1494 patients with appropriate International Classification of Diseases codes. From this list, 620 patient slides were selected at random and scanned. After scanning, slide quality was judged to be poor because of fading in 78 slides, and these were excluded from further analysis. Therefore, we included 542 patients in our study: 164 patients with NDBE (30.3%), 226 with LGD (41.7%), and 152 with HGD (28.0%). Basic demographics are shown in Table 1.

The training set consisted of 368 slides (67.9%), the validation set 104 slides (19.2%), and the test set 70 slides (12.9%). This yielded 2380 unique bounding boxes around the previously manually drawn annotations. For the object detection model (requiring conversion of these images to 1280 × 1280-pixel tiles), we analyzed 11,382 bounding boxes (8596, 1946, and 840 for training, validation, and test sets, respectively) (Fig. 1E). For the classifier model (requiring conversion to 1024 × 1024-pixel tiles), we analyzed 12,196 unique bounding boxes yielding 8914, 2248, and 1034 bounding boxes for training, validation, and test sets, respectively (Fig. 2B).

Interobserver agreement of study pathologists

After review of a random sample of 70 slides (from the 542-slide cohort), the overall agreement between study pathologists was substantial, with a kappa statistic of .72 (95% confidence interval, .59-.86).

Accuracy of the deep learning model in diagnosing BE dysplasia

WSI-level ensemble performance. To determine model performance when analyzing WSIs, we converted WSIs of the test set (n = 70) into 1280 × 1280-pixel tiles, resulting in 394,396 tiles. We subsequently fed the regions detected as abnormal by the YOLO model into the classifier model for ensemble analysis. The combined system achieved 88.6% accuracy for 3-class distinction (NDBE vs LGD vs HGD) (Table 2) and 95.7% accuracy for 2-class distinction (NDBE vs dysplastic BE) (Table 2).

Table 3 demonstrates the confusion matrix of the WSI-level ensemble approach for the 3-class discrimination. Of note, the model misclassified 2 cases of LGD as NDBE but otherwise did not underdiagnose any other slides. Only 1 patient with NDBE was misdiagnosed by the ensemble model as HGD.

This "2-step" model identified 20 of 5996 tiles that had higher-grade findings than originally annotated by study pathologists. After additional review by study pathologists, 5 of these 20 upgraded predictions were confirmed to be correct. Individual model performance (object detection [YOLO] and classifier models alone) results are presented in the Supplementary Methods.

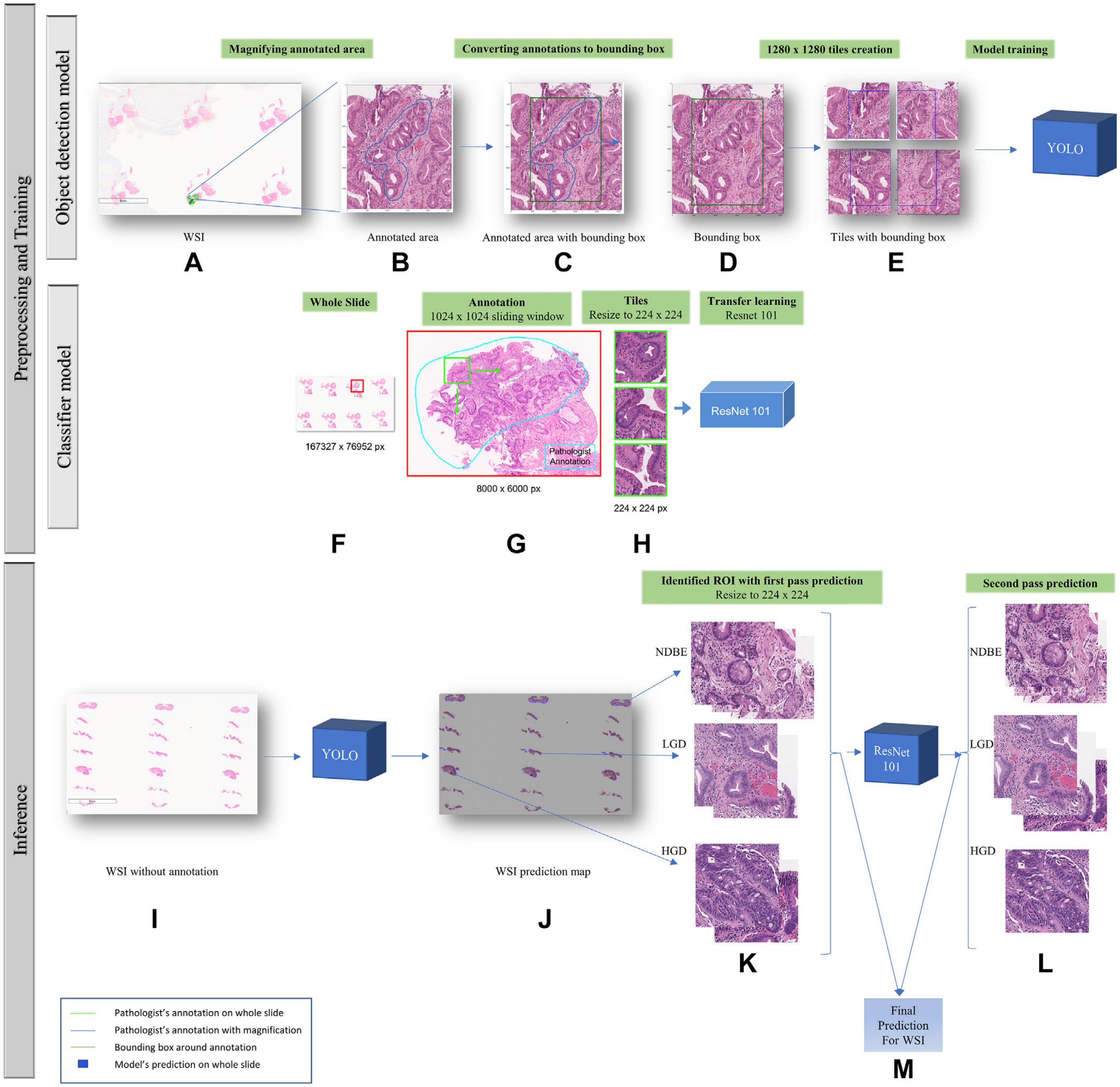
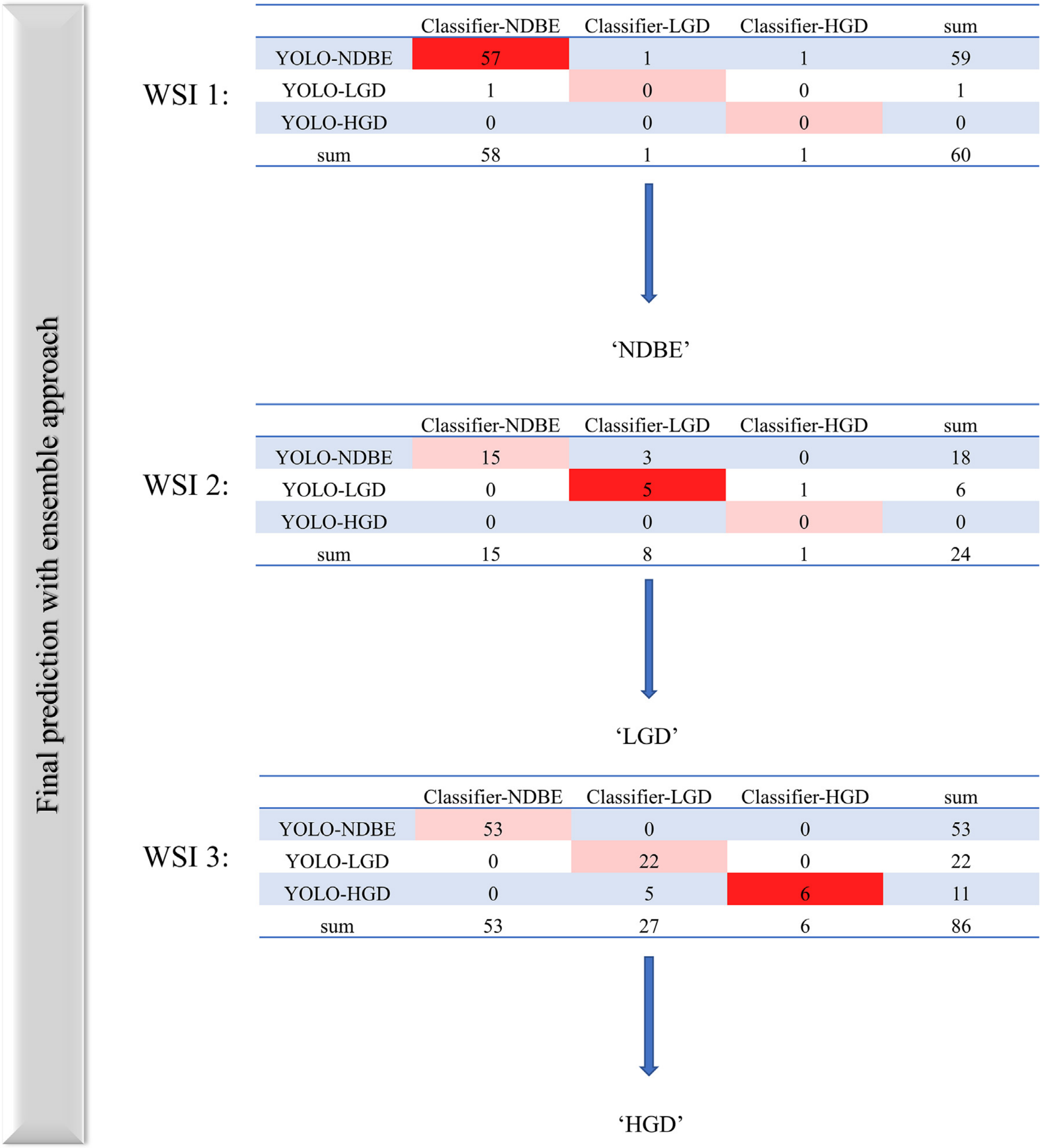


Figure 1. Summary of preprocessing and training of object detection and classifier models and the process for inference by ensemble model. **A-E,** Object detection model development. **F-H,** Classifier model development. **I-M,** Ensemble inference. **A,** Whole-slide images demonstrating areas of annotations by expert GI pathologists. **B,** Magnification view of annotation. **C,** Creation of bounding box around annotation. **D,** Segmentation of bounding box into 1280 × 1280 tile components. **E,** Training for object detection (YOLO) model. **F,** Whole slide with area of annotation. **G,** Close-up view of annotated areas. **H,** Resizing of annotations to 224 × 224 tile segments followed by training of classifier model. **I,** Removal of pathology annotations from test set slides with first-pass object detection model assessment. **J,** Regions of interest as identified by the object detection model. **K,** Resizing of identified regions of interest to 224 × 224 pixels for feeding into object detection mode. **L,** Second-pass prediction with the classifier model. **M,** Final dysplasia grade prediction by ensemble model. *YOLO*, You only look once. *WSI*, Whole slide imaging; *px*, pixels; *NDBE*, non-dysplastic Barrett's esophagus; *LGD*, low-grade dysplasia; *HGD*, high- grade dysplasia; *ROI*, regions of interest.

WSI prediction map. Figure 1J shows the ensemble model prediction map. Of note, the model-predicted dysplasia areas appear well aligned with the original pathologists' annotations.

DISCUSSION

We demonstrated the successful development and validation of an ensemble deep learning model using an



WSI 1:

	Classifier-NDBE	Classifier-LGD	Classifier-HGD	sum
YOLO-NDBE	57	1	1	59
YOLO-LGD	1	0	0	1
YOLO-HGD	0	0	0	0
sum	58	1	1	60

WSI 2:

	Classifier-NDBE	Classifier-LGD	Classifier-HGD	sum
YOLO-NDBE	15	3	0	18
YOLO-LGD	0	5	1	6
YOLO-HGD	0	0	0	0
sum	15	8	1	24

WSI 3:

	Classifier-NDBE	Classifier-LGD	Classifier-HGD	sum
YOLO-NDBE	53	0	0	53
YOLO-LGD	0	22	0	22
YOLO-HGD	0	5	6	11
sum	53	27	6	86

Figure 2. Process of ensemble model histologic dysplasia grade prediction. The *first column* demonstrates the object detection (YOLO) predictions, and the *first row* displays the classifier model’s predictions. In the matrix, the diagonal axis shows the agreement between the object detection and classifier models. We chose the highest grade of dysplasia as agreed on by both models as the final whole-slide histologic prediction. *YOLO*, You only look once; *WSI*, whole-slide image; *NDBE*, nondysplastic Barrett’s esophagus; *LGD*, low-grade dysplasia; *HGD*, high-grade dysplasia.

object detection (YOLOv5) model for efficient WSI analysis combined with a classifier model adept in image and class discrimination tasks to accurately predict dysplasia grade on digital hematoxylin and eosin BE slide images. We report high sensitivity and specificity as well as high accuracy, using agreement between 2 expert pathologists on dysplasia grade as the criterion standard.

Our results build on prior work done on artificial intelligence–enhanced histologic analysis of BE slides. Tomita et al[11] analyzed 379 histologic images obtained from 180 slides using an attention-based deep neural network. They reported sensitivities ranging from 21% to 77% for the diagnosis of squamous tissue, NDBE, dysplastic BE, and EAC. The greater number of whole slides analyzed in our training and validation sets (472 unique slides) and the use of a novel ensemble methodology likely explain the improved performance noted in our model.

The difficulty in diagnosing LGD has led to controversy in the management of LGD.[6] Prior work has

TABLE 1. Baseline demographics of patients included in model development and validation

	Training set (n = 368)	Validation set (n = 104)	Test set (n = 70)	*P* value	Total
Mean age, y (SD)	64.2 (11.1)	63.9 (9.8)	62.5 (9.8)	>.05	62.4 (13.7)
Male gender, %	81.5	76.6	79.2	>.05	71.4
Mean BE length, cm (SD)	5.3 (3.5)	5.4 (4.0)	6.0 (4.1)	>.05	
Nondysplastic BE	111 (30.1)	31 (29.8)	18 (25.7)		160 (30.3)
Low-grade dysplasia	153 (41.6)	43 (41.3)	33 (47.1)		229 (41.7)
High-grade dysplasia	104 (28.3)	30 (28.8)	19 (27.1)		153 (28.0)

Values are n (%) unless otherwise defined.
BE, Barrett's esophagus.

TABLE 2. Ensemble model discrimination among 3 classes (NDBE vs low-grade dysplasia vs high-grade dysplasia) and among 2 classes (NDBE vs dysplastic BE) on whole-slide imaging

Class	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)	F1 score
3-class model					
NDBE (n = 18)	94.4 (72.7-99.8)	96.2 (86.7-99.5)	99.2 (97.2-99.8)	75.3 (31.2-95.3)	.919
Low-grade dysplasia (n = 33)	81.8. (64.5-93)	97.3 (85.8-99.9)	77 (32.5-95.9)	97.9 (95.9-99)	.885
High-grade dysplasia (n = 19)	94.7 (73.9-99.8)	90.2 (78.5-96.7)	33.7 (18-54.9)	99.6 (97.9-99.9)	.857
2-class model					
NDBE (n = 18)	94.4 (72.7-99.8)	96.2 (86.7-99.5)	99.2 (97.2- 99.8)	75.3 (31.2-95.3)	.919
Dysplastic BE (n = 52)	96.2 (86.7-99.5)	94.4 (72.7-99.8)	75.3 (31.2-95.3)	99.2 (97.2-99.2)	.971

Values in parentheses are 95% confidence intervals.
NDBE, Nondysplastic Barrett's esophagus; *BE*, Barrett's esophagus.

TABLE 3. Confusion matrix for the ensemble model test set by whole-slide images (3-class system)

Actual	Nondysplastic Barrett's esophagus	17	0	1
	Low-grade dysplasia	2	27	4
	High-grade dysplasia	0	1	18
		Nondysplastic Barrett's esophagus	Low-grade dysplasia	High-grade dysplasia
		Predicted		

demonstrated that pathologists are able to discriminate NDBE/indefinite for dysplasia/LGD from HGD/EAC quite easily, but when assessing individual dysplastic states of BE, consensus falters.[23] More-recent studies have demonstrated persistent discordance in the diagnosis of LGD among pathologists, even in those considered "experienced."[24] Indeed, interobserver variation, even among academic pathologists, is quite high, with kappa statistics frequently <45%.[25] Additionally, up to 85% of community-diagnosed LGD is downgraded on review of histology by pathologists with expertise in BE histology.[6] The importance of reliably diagnosing LGD is supported by studies demonstrating that when multiple pathologists agree on the diagnosis, the risk of developing EAC increases substantially.[6,26]

The current critical need related to this diagnostic uncertainty stems from our consequent inability to accurately guide treatment in this cohort of patients. Although confirmed LGD carries an elevated progression risk,[27] variability in progression estimates[28] has led to recommendations that suggest either ablation for persistent disease or continued surveillance until progression is noted using shared decision-making.[3,29] Neither of these options are without risks because ablation can be complicated by periprocedural pain and adverse events requiring procedural management, whereas observation alone may lead to a delayed diagnosis of invasive cancer.

Hence, confirmation by an expert BE pathologist is recommended for community-diagnosed LGD to reduce the risk of overly aggressive therapy or unwarranted observation. However, the associated cost, limited availability of "expert pathologists," and poor interobserver agreement even between expert pathologists all limit the utility of this approach. Application of artificial intelligence and deep learning models may mitigate these challenges, and with increased certainty of diagnosis, management may become more efficient, selecting patients who may truly benefit from EET. Avoidance of invasive EET in patients when it is not indicated is also equally important.

Consensus suggests that larger amounts of data improve the performance of image-analysis tasks. We acknowledge that training on a training set size that is not ideal may increase the risk of overfitting, and this is a limitation because of the relative paucity of dysplastic BE slides compared with the abundance of slides or images available in other similar analysis tasks (eg, polyp detection models) and the relatively small percentage of an esophageal biopsy sample slide, which may harbor dysplastic tissue. To mitigate this, we used various methods of data augmentation, used early stopping during model development, and trained with heterogeneous slides obtained during a broad timeframe of clinical practice. We also used a hold-out test set to assess model efficacy, which can avoid any bias that may be introduced by training and testing on the same images.

Our study has several strengths. We used a novel, ensemble, 2-step algorithm consisting of 2 separate but integrated deep learning models adept at object detection and classification tasks to improve the accuracy of diagnosing dysplastic BE based on WSI analysis. The use of an object detection model to make first-pass predictions significantly improves the efficiency of the process; otherwise, significant computational power would be required for a classifier model to assess the millions of pixels that make up a histologic slide and given that dysplastic BE typically involves a very small percentage of an overall slide. We have drawn from a large sample of patients with confirmed endoscopic and histologic BE, and regions with dysplasia were annotated by expert BE pathologists with substantial experience in the diagnosis of dysplastic BE. This strengthens the quality of the slides' diagnoses on which our model was trained. We were able to train, validate, and independently test our results on several thousand images, strengthening the model heuristics. We also developed prediction plots that provide an element of explainability and strengthen our results.

There are some potential limitations to be noted. Given the timespan of this study, the histopathologic definition of dysplasia may have changed over time, but re-review by an expert GI study pathologist in the present time should mitigate any bias this may introduce. We did not use normal squamous tissue or EAC in our training image repository, given that diagnostic uncertainty in these histologic categories (no intestinal metaplasia [IM] vs IM and EAC vs no EAC) is believed to be lower. Although the confusion matrix did show some misplaced results, this is still a considerable improvement compared with interobserver agreement among expert pathologists in prior studies.[6,24] In the ensemble analysis, only 3 cases of LGD were marked as NDBE. This must be minimized going forward because patients could miss therapeutic interventions. Future research will include external validation in larger multicenter image repositories, inclusion of all biopsy samples and endoscopic resections obtained from a procedure, and refinements of the model heuristics with a focus on improved LGD diagnosis sensitivity. However, it is important to recognize that overall the model was tested against expert GI pathologists with considerable experience in Barrett's neoplasia. As such, its sensitivity against community-diagnosed LGD is likely to be significantly higher.

In conclusion, we demonstrated the development and internal validation of a novel, 2-step, deep learning model to identify dysplasia grades on digitized histology slides. External validation is the next logical step in development of this model and is ongoing. This will hopefully enable more accurate targeting of EET for the treatment of dysplastic BE.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin 2020;70:7-30.
2. Rastogi A, Puli S, El-Serag HB, et al. Incidence of esophageal adenocarcinoma in patients with Barrett's esophagus and high-grade dysplasia: a meta-analysis. Gastrointest Endosc 2008;67:394-8.
3. Shaheen NJ, Falk GW, Iyer PG, et al. ACG clinical guideline: diagnosis and management of Barrett's esophagus. Am J Gastroenterol 2016;111:30-50; quiz 51.
4. Qumseya B, Sultan S, Bain P, et al. ASGE guideline on screening and surveillance of Barrett's esophagus. Gastrointest Endosc 2019;90:335-59.
5. Odze RD. Diagnosis and grading of dysplasia in Barrett's oesophagus. J Clin Pathol 2006;59:1029-38.
6. Curvers WL, ten Kate FJ, Krishnadath KK, et al. Low-grade dysplasia in Barrett's esophagus: overdiagnosed and underestimated. Am J Gastroenterol 2010;105:1523-30.
7. Huang G, Liu Z, Maaten LVD, et al. Densely connected convolutional networks. Presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017. Available at: https://ieeexplore.ieee.org/document/8099726. Accessed July 21, 2022.
8. Anuse A, Vyas V. A novel training algorithm for convolutional neural network. Complex Intell Syst 2016;2:221-34.
9. Aoki T, Yamada A, Aoyama K, et al. Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network. Gastrointest Endosc 2019;89:357-63.
10. Marya NB, Powers PD, Chari ST, et al. Utilisation of artificial intelligence for the development of an EUS-convolutional neural network model trained to enhance the diagnosis of autoimmune pancreatitis. Gut 2021;70:1335-44.
11. Tomita N, Abdollahi B, Wei J, et al. Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. JAMA Netw Open 2019;2:e1914645.
12. Beuque M, Martin-Lorenzo M, Balluff B, et al. Machine learning for grading and prognosis of esophageal dysplasia using mass spectrometry and histological imaging. Comput Biol Med 2021;138:104918.
13. Hansen LK, Salamon P. Neural network ensembles. IEEE Trans Pattern Anal Machine Intell 1990;12:993-1001.

14. Goldblum JR. Current issues in Barrett's esophagus and Barrett's-related dysplasia. Mod Pathol 2015;28(Suppl 1):S1-6.
15. Naini BV, Souza RF, Odze RD. Barrett's esophagus: a comprehensive and contemporary review for pathologists. Am J Surg Pathol 2016;40:e45-66.
16. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Machine Learn Res 2011;12:2825-30.
17. Ganaie MA, Hu M, Tanveer M, et al. Ensemble deep learning: a review. ArXiv 2021;abs/2104.02395.
18. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition: IEEE Computer Society; 2014. p. 580-7.
19. Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection. ArXiv 2015;1506:02640.
20. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. ArXiv 2014:1409-4842.
21. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems 1. Lake Tahoe, NV: Curran Associates Inc; 2012. p. 1097-105.
22. Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE 1998;86:2278-324.
23. Montgomery E, Bronner MP, Goldblum JR, et al. Reproducibility of the diagnosis of dysplasia in Barrett esophagus: a reaffirmation. Hum Pathol 2001;32:368-78.
24. Vennalaganti P, Kanakadandi V, Goldblum JR, et al. Discordance among pathologists in the United States and Europe in diagnosis of low-grade dysplasia for patients with Barrett's esophagus. Gastroenterology 2017;152:564-70.
25. Alikhan M, Rex D, Khan A, et al. Variable pathologic interpretation of columnar lined esophagus by general pathologists in community practice. Gastrointest Endosc 1999;50:23-6.
26. Krishnamoorthi R, Lewis JT, Krishna M, et al. Predictors of progression in Barrett's esophagus with low-grade dysplasia: results from a multicenter prospective BE registry. Am J Gastroenterol 2017;112:867-73.
27. Duits LC, van der Wel MJ, Cotton CC, et al. Patients with Barrett's esophagus and confirmed persistent low-grade dysplasia are at increased risk for progression to neoplasia. Gastroenterology 2017;152:993-1001.
28. Singh S, Manickam P, Amin AV, et al. Incidence of esophageal adenocarcinoma in Barrett's esophagus with low-grade dysplasia: a systematic review and meta-analysis. Gastrointest Endosc 2014;79:897-909.e4.
29. Krishnamoorthi R, Hargraves I, Gopalakrishnan N, et al. Development and pilot testing of decision aid for shared decision making in Barrett's esophagus with low-grade dysplasia. J Clin Gastroenterol 2021;55:36-42.

Abbreviations: BE, Barrett's esophagus; EAC, esophageal adenocarcinoma; EET, endoscopic eradication therapy; NDBE, nondysplastic Barrett's esophagus; HGD, high-grade dysplasia; LGD, low-grade dysplasia; WSI, whole-slide image; YOLO, you only look once.

DISCLOSURE: *The following authors disclosed financial relationships: B. J. Erickson: Officer at FlowSIGMA, Inc and Yunu, Inc. P. G. Iyer: Research funding from Exact Sciences, Pentax Medical, and Cernostics; consultant for Exact Sciences, Medtronic, Ambu Medical, Symple Surgical, and Cernostics. C. L. Leggett: research support for Verily Life. All other authors disclosed no financial relationships. Research support for this study was provided by Mayo Clinic Division of Gastroenterology and Hepatology Innovation Award and the Freeman Foundation.*

DIVERSITY, EQUITY, AND INCLUSION: One or more of the authors of this paper self-identifies as an under-represented gender minority in science. While citing references scientifically relevant for this work, we actively worked to promote gender balance in our reference list.

*Drs Faghani and Codipilly and Erickson and Iyer contributed equally to this article.

0016-5107/$36.00
https://doi.org/10.1016/j.gie.2022.06.013

Received April 19, 2022. Accepted June 8, 2022.

Current affiliations: Artificial Intelligence Laboratory (1), Department of Radiology (2), Barrett's Esophagus Unit, Division of Gastroenterology and Hepatology (3), Department of Pathology, Mayo Clinic, Rochester, Minnesota, USA (4); Department of Pathology, Mayo Clinic, Jacksonville, Florida, USA (5).

Reprint requests: Prasad G. Iyer, MD, MSc, Barrett's Esophagus Unit, Division of Gastroenterology and Hepatology, 200 1st St SW, Mayo Clinic, Rochester, MN 55905.

SUPPLEMENTARY METHODS

Object detection (you only look once) model development

Data preprocessing. Magnification levels for whole-slide images (WSIs) were set at 40× (Fig. 1B). We aligned segmented annotations to the bounding box coordinates on each slide. We chose the minimum x and y coordinates of each segmented area as the x and y coordinates of the top-left point of the bounding box accordingly and the maximum x and y coordinates as the bottom-right point (Fig. 1C). Because of high resolution, annotated areas could not be used directly as inputs into the current deep learning models. To address this challenge, we converted each annotated area along with their corresponding bounding boxes into nonoverlapping tiles of 1280 × 1280 pixels (which was the input size of the you only look once [YOLO] object detection model) (Fig. 1E).

Model development. We have previously trained several state-of-the-art object detection models, including MMdetection and different models of YOLO with and without transfer learning.[1-4] Ultimately, we used a YOLOv5m6 pretrained on the common object in context dataset[5] because of the superior performance on prior validation set testing and greater efficiency during inference. We optimized the model hyperparameters using a grid search (Supplementary Table 1). We trained the model for 100 epochs with early stopping criteria using horizontal and vertical flipping; scaling; translating; changing hue, saturation, and value; and mosaics for data augmentation, which helps to avoid overfitting (Supplementary Table 1). We selected the model weights with the highest mean average precision, a common performance metric for object detection with .5 as the threshold for the intersection over union (mean average precision, .5) on the validation set.

We converted each WSI into 1280 × 1280-pixel nonoverlapping tiles before inference. For inference on the test set, we used test time augmentation, an intersection over union threshold of .3, and a confidence level of .7, which was determined through optimization on the validation set. Using our object detection model, we identified the regions containing the pathology of interest on each of the tiles. Mean average precision of .5, precision (positive predictive value), recall (sensitivity), specificity, and harmonic mean of precision and recall (F1 score) were calculated for each of the classes at the tile level. The F1 score is an important metric in analyzing multiclass results because it considers missed cases, with particular weight assigned to extreme (clinically impactful in this study) results, which are important to factor because missed high-grade dysplasia (HGD) may have poor long-term consequences.

Prediction map. We created a prediction map (Fig. 1J) for each slide to see where the model predicted areas of dysplasia on the slides. Each grade was marked with a different color (blue for nondysplastic Barrett's esophagus [NDBE], yellow for low-grade dysplasia [LGD], and red for HGD). We used the prediction map as a visual check to assess if the model was detecting pathology in the correct areas of the slides.

Classifier model development

Data preprocessing. At 40× magnification, we converted annotated areas into 1024 × 1024 tiles (Fig. 1G), ensuring that at least 80% of the tile would be inside the annotated area images, and then resized tiles to 224 × 224 pixels (input size of the classifier model) (Fig. 1H).

Model development. We initially evaluated several different deep learning architectures, including ResNet 18, ResNet 34, ResNet 50 (without and with ImageNet pretraining), ResNet 101, ResNet 152 (without and with ImageNet pretraining), DenseNet 121, and DenseNet 201.[6] We found that a ResNet101 model, pretrained on ImageNet[7], from the Fastai computer vision models library performed best during preliminary testing and subsequently used only this architecture.[8,9] The 3-class model was trained to distinguish between NDBE, LGD, and HGD. To mitigate the mildly imbalanced training dataset, the under-represented classes were oversampled (in a 3:3:1 NDBE/LGD/HGD ratio), and to avoid overfitting, the training data were augmented using vertical flip, rotate, zoom, lighting, warp, and random erasing. Again, pathologist annotations were removed from the test set slides. Fastai's Hyperparam schedule fine_tune was used for model training. The optimizer was ADAM, and the loss function was categorical cross-entropy. The last 2 layers of the model were trained for 5 epochs with a base learning rate of 2e-3, whereas the other layers were "frozen" (weights not changed). All layers of the model were then unfrozen with a base learning rate of 2e-3 for 15 epochs. For a second time the last 2 layers of the model were trained for 5 epochs with a base learning of 1e-6. Again, all layers of the model were then trained with a base learning rate of 1e-6, and the training was stopped after 15 epochs when the model accuracy did not improve on the validation set.

A 2-class model (dysplasia vs no dysplasia) was also trained using a pretrained ResNet 101. From a clinical standpoint, the 2-class model is helpful because it can help "rule out" patients with NDBE who would not require endoscopic eradication therapy such as ablation or resection for management. The training data were augmented using vertical flip, rotate, zoom, lighting, warp, and random erasing. Fastai's Hyperparam schedule fine_tune was used for training the model. The hyperparameters and training schema was the same as for the 3-class model.

The 3-class model was used to predict the highest class (NDBE, LGD, or HGD) of each tile in the test set. We then collected the classes predicted for all tiles from a patient and used a majority vote schema to predict the class for that patient: A patient was classified as NDBE, LGD, or HGD based on the majority tile prediction by the model.

Sensitivity, specificity, and the harmonic mean of precision and recall were calculated for each of the 3 classes. A confusion matrix was also created to assess model mischaracterization, allowing for insight into whether the model was over- or under-diagnosing classes. This was also done for the 2-class model, except of course only 2 classes (dysplasia or nondysplastic) were considered.

REFERENCES

1. Mmdetection. OpenMMLab detection toolbox and benchmark. Available at: https://github.com/open-mmlab/mmdetection. Accessed December 21, 2021.
2. Chen K, Wang J, Pang J, et al. Open mmlab detection toolbox and benchmark. Available at: https://arxiv.org/abs/1906.07155. Accessed July 21, 2022.
3. Liu X, Li M, Hao F, et al. GLO-YOLO: a dynamic glomerular detecting and slicing model in whole slide images. In: Proceedings of the 2020 Conference on Artificial Intelligence and Healthcare. Taiyuan, China: Association for Computing Machinery; 2020. p. 229-33.
4. Swiderska-Chadaj Z, Pinckaers H, van Rijthoven M, et al. Learning to detect lymphocytes in immunohistochemistry with deep learning. Med Image Anal 2019;58:101547.
5. Lin T-Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. In: Computer Vision—ECCV 2014. Cham, Switzerland: Springer International Publishing; 2014.
6. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. ArXiv 2015;1512:03385.
7. Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. Available at: https://ieeexplore.ieee.org/document/5206848. Accessed July 21, 2022.
8. Howard J, Gugger S. Fastai. A layered API for deep learning. Information 2020;11:108.
9. Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks. ArXiv 2016;1608:06993.

SUPPLEMENTARY TABLE 1. Optimized object detection model hyperparameters

Hyperparameter	Value
Image size	1280 × 1280 pixels
Batch size	8
Optimizer	Stochastic gradient descent
Initial learning rate	.01
Final one cycle learning rate	.2
Momentum	.937
Weight decay	.0005
Warmup epochs	3
Warmup momentum	.8
Warmup bias learning rate	.1
Image HSV: hue augmentation (fraction)	.015
Image HSV: saturation augmentation (fraction)	.7
Image HSV: value augmentation (fraction)	.4
Image translation (fraction)	.1
Image flip up-down (probability)	.5
Image flip left-right (probability)	.5
Image scale (± gain)	.9
Image mosaic (probability)	1

HSV, Hue, saturation, value.